Bayesian Regularisation in Geoadditive Expectile Regression

Elisabeth Waldmann, Fabian Sobotka, Thomas Kneib

April 20, 2015

Abstract

Regression classes modeling more than the mean of the response have found a lot of attention in the last years. Expectile regression is a special and computationally convenient case of this family of models. Expectiles offer a quantile-like characterisation of a complete distribution and include the mean as a special case. In the frequentist framework the impact of a lot of covariates with very different structures have been made possible. We propose Bayesian expectile regression based on the asymmetric normal distribution. This renders possible incorporating for example linear, nonlinear, spatial and random effects in one model as well as Bayesian regularization. Furthermore a detailed inference on the estimated parameters can be conducted. Proposal densities based on iteratively weighted least squares updates for the resulting Markov chain Monte Carlo (MCMC) simulation algorithm are proposed and the potential of the approach for extending the flexibility of expectile regression towards Spike-and-Slab regularization as well as complex semiparametric regression specifications is discussed.

Expectile Regression, Bayesian Semiparametric Regression, Markov random fields, P-splines, asymmetric normal distribution, Markov chain Monte Carlo Simulation, Spike and Slab priors

1 Introduction

Recent interest in the development of flexible regression specifications has had a specific focus on describing more complex features of the response

distribution than only the mean. The standard instrument in this situation is quantile regression (Koenker and Bassett, 1978) where conditional quantiles are related to a regression predictor. A lot of work has been done to extend the simple linear quantile regression model to more advanced approaches like quantile smoothing splines (Koenker et al., 1994), quantile regression for clustered data (Reich et al., 2010) or geoadditive models (Fenske et al., 2011).

Computationally regression quantiles are obtained by minimising an asymmetrically weighted absolute residuals criterion

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})|y_i - \eta_{i\tau}| \tag{1}$$

with asymmetric weights

$$w_\tau(y_i, \eta_{i\tau}) = \begin{cases} 1-\tau & y_i \leq \eta_{i\tau} \\ \tau & y_i > \eta_{i\tau}, \end{cases}$$

a response y and a quantile-specific predictor η_τ. This loss function induces additional complexity compared to standard least squares optimisation. As a consequence, expectile regression (Newey and Powell, 1987) that relies on asymmetrically weighted squared residuals

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})(y_i - \eta_{i\tau})^2 \tag{2}$$

has gained considerable interest since expectile regression estimates can be obtained by simple iteratively weighted least squares fits. Extensions to more complicated models have been explored in recent publications for the smoothing of a nonlinear effect (Schnabel and Eilers, 2009), for geoadditive models (Sobotka and Kneib, 2012) and for instrumental variables (Sobotka et al., 2013). While basic asymptotic results are available for a least squares estimate (see Sobotka et al., 2013), alternative estimation methods like boosting as introduced to expectiles by Sobotka and Kneib (2012) rely on a bootstrap for further inference. An autoregressive definition of expectiles was even introduced for time series analysis (Taylor, 2008). In this paper, we introduce a Bayesian formulation of expectile regression that relies on the asymmetric normal distribution (AND) as auxiliary response distribution. The approach

is very similar to the estimation of Bayesian quantile regression, where an asymmetric Laplace distribution (ALD) is used instead of the AND. For detailed information see Yue and Rue (2011), Kozumi and Kobayashi (2011) or Reed and Yu (2009). In the case of the AND proposal densities based on iteratively weighted least squares updates for the resulting Markov chain Monte Carlo (MCMC) simulation algorithm are needed.

As an illustrative example, we present a data set dealing with malnutrition in Tanzania. The dependent variable is the so called *z-score* of *stunting* (a score measuring the height of the child in comparison to a reference population). The latter is the dependent variable and shall be explained by continous covariates like *maternal BMI at birth*, *age of the child* and categorical covariates (*mother's work*, *mother's education* and *mother's residence*, denoted by $\boldsymbol{X}$). The impact of the continuous covariates used for the explanation of the dependent variable *z-score* is not linear thus we use splines. As Tanzania consists of 20 regions over which economic and political situation differ we will also to incorporate the regions into the model. Therefore use a geoadditive model of the type

$$\text{stunting}_i = f(\text{BMI}_i) + f(\text{age}_i) + f_{geo}(\text{region}_i) + \boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i, \tag{3}$$

where the f denotes the nonlinear effects, $\boldsymbol{X}$ contains categorical covariates, which will be se and f_{geo} is the spatial effect of the different regions of Tanzania. The necessity of using a model different to mean regression becomes obvious when taking a look at the data: the conditional distribution of the z-score is neither homoscedastic nor symmetric.

The rest of the paper is structured as follows: in the second section we describe the basic ideas of expectile regression and give an overview over the concept of semiparametric regression and variable selection via Spike-and-Slab priors. We then introduce the above mentioned asymmetric normal distribution and describe the Bayesian algorithm in more detail. The third section contains simulations which study point estimation as well as confidence intervals for the parameters in the first part. The second set of simulations aims to evaluate the performance of the regularization by the Spike-and-Slab priors. In Section 4 we will describe the above mentioned data set on childhood malnutrition in Tanzania and explain the impact of the different covariates. In the last section we conclude and give an outlook on future plans.

2 Bayesian Expectile Regression

2.1 Expectile Regression

Suppose that regression data $(y_i, \boldsymbol{z}_i)$, $i = 1, \ldots, n$, on a continuous response variable y and a covariate vector $\boldsymbol{z}$ are given and shall be analysed in a regression model of the form

$$y_i = \eta_{i\tau} + \varepsilon_{i\tau},$$

where η_τ is a predictor formed by the covariates and ε_τ is an appropriate error term. Unlike in mean regression where regression effects on the mean are of interest, we focus on situations where specific outer parts of the response distribution shall be studied. We will denote the extremeness of these outer parts by the asymmetry parameter $\tau \in (0, 1)$ where $\tau = 0.5$ corresponds to the central part of the distribution while $\tau \to 0$ and $\tau \to 1$ yield the lower and upper part of the distribution, respectively. The standard approach for implementing such regression models is quantile regression where we assume that the τ-quantile of the error distribution equals zero, i.e.

$$P(\varepsilon_{i\tau} \leq 0) = \tau.$$

This implies that the predictor $\eta_{i\tau}$ corresponds to the τ-quantile of the response y_i and the regression model can be estimated by minimising the loss function (1). As an alternative, we will instead focus on the criterion (2) that yields expectile regression estimates. This criterion has the advantage to be differentiable with respect to the regression predictor so that estimates can be obtained by iteratively weighted least squares estimation. Basically, expectiles are an alternative possibility to characterise the distribution of a continuous random variable where τ indicates the "extremeness" of the part of the distribution that shall be studied, see Newey and Powell (1987).

A usual objection against expectiles as compared to quantiles is their lack of an immediate interpretation. While for quantiles the property that $\tau \cdot 100$ percent of the data lie below the regression line and $(1 - \tau)100$ percent of the data lie above the regression line is easy to understand, the extremeness of expectiles is hard to transfer to such an easy statement. However, interpretation of expectiles is still possible in the following ways:

- For i.i.d. data $y_1, \ldots, y_n$, the resulting expectile estimate $\hat{e}_\tau$ will be a

weighted average

$$\hat{e}_\tau = \sum_{i=1}^{n} w_i y_i$$

where the weights w_i depend on the estimated expectile. As a consequence, regression expectiles can also be considered such a weighted average conditioned on a specific covariate vector.

- Expectiles are tail expectations, i.e. the τ-expectile fulfills

$$\tau = \frac{\int_{-\infty}^{e_\tau} |y - e_\tau| f(y) dy}{\int_{-\infty}^{\infty} |y - e_\tau| f(y) dy}$$

showing that e_τ is characterised by a partial moment condition.

- Usually, one would not only estimate one single expectile but a whole set of expectiles for various values of τ. The collection of all estimates then gives an intuitive impression about the shape of the conditional distribution of the response and in particular allows to detect features such as heteroscedasticity, skewness or kurtosis. Moreover, conditional quantiles can still be calculated from a set of expectiles if quantile estimates are of ultimate interest, as shown by Efron (1991) and refined in Schulze Waltrup et al. (2013).

- Expectiles are increasingly important when it comes to measuring risks. Taylor (2008) uses expectiles to efficiently estimate the expected shortfall (ES), a coherent and subadditive risk measure. Its estimation would normally base on a small subset of the available sample. In contrast, the estimate based on expectiles contains all observations. Recent results by Ziegler (2013) also show that expectiles themselves are a coherent and elicitable risk measure while quantiles are not coherent.

In summary, albeit having a different (and may be less intuitive) interpretation than quantiles, expectiles are probably not more difficult to interpret than a variance.

2.2 Asymmetric Normal Distribution

To make expectile regression accessible in a Bayesian formulation, we require the specification of an auxiliary response distribution that yields a likelihood

that is equivalent to the optimisation criterion (2). For Bayesian quantile regression, this can be formalised based on the asymmetric Laplace distribution, see for example Yue and Rue (2011), Lum and Gelfand (2012) or Yu and Moyeed (2001). For expectile regression, the analogous distribution is an asymmetric normal distribution

$$y_i \sim \mathrm{AN}(\eta_i, \sigma^2, \tau)$$

with density

$$p(y_i) = \frac{2}{\sqrt{\sigma^2\pi}} \left(\sqrt{\frac{1}{1-\tau}} + \sqrt{\frac{1}{\tau}} \right)^{-1} \exp\left(-\frac{1}{\sigma^2} \omega_\tau(y_i, \eta_{i\tau})(y_i - \eta_{i\tau})^2 \right).$$

expectation

$$\mathrm{E}(y_i) = \eta_{i,\tau} + \frac{\sigma}{(\sqrt{\tau} + \sqrt{1-\tau})} \left(\frac{1-2\tau}{\sqrt{\pi\tau(1-\tau)}} \right)$$

and variance

$$\mathrm{Var}(y_i) = \frac{\sigma^2}{\sqrt{\tau} + \sqrt{1-\tau}} \left[\frac{1}{2} \left(\frac{\sqrt{1-\tau}}{\tau} + \frac{\sqrt{\tau}}{1-\tau} \right) - \frac{1}{\sqrt{\tau} + \sqrt{1-\tau}} \left(\frac{(1-2\tau)^2}{\pi\tau(1-\tau)} \right) \right].$$

Maximising the likelihood arising from this distributional specification is then equivalent to minimising (2), as the logarithmic kernel of the distribution is the same (but negative) argument.

2.3 Semiparametric Regression

Instead of only considering linear regression specifications, we are interested in applying expectile regression in the context of general semiparametric regression models with predictor

$$\eta_i = \beta_0 + \sum_{j=1}^{p} f_j(\boldsymbol{z}_i),$$

where β_0 is an intercept representing the overall level of the predictor, and the functions $f_j(\boldsymbol{z}_i)$ reflect different types of regression effects depending on subsets of the covariate vector $\boldsymbol{z}_i$. Note that we suppress the index τ for notational simplicity. For the regression functions f_j, we make the following assumptions:

- The functions f_j are approximated in terms of basis function representations
$$f_j(\boldsymbol{z}) = \sum_{k=1}^{K} \beta_{jk} B_k(\boldsymbol{z})$$
where $B_k(\boldsymbol{z})$ are the basis functions and β_{jk} denote the corresponding basis coefficients.

- The conditional prior for the vector of basis coefficients $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})'$ given hyperparameters θ_j is a multivariate normal distribution with density
$$p(\boldsymbol{\beta}_j|\boldsymbol{\theta}_j) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}_j' \boldsymbol{K}_j(\boldsymbol{\theta}_j)\boldsymbol{\beta}_j\right)$$
where the precision matrix $\boldsymbol{K}_j(\boldsymbol{\theta}_j)$ either represents different types of structural assumptions about the function f_j such as smoothness or induces regularisation to a set of covariates. The vector $\boldsymbol{\theta}_j$ can contain potential hyperparameters which can for example control shrinkage or the degree of smoothness, as will be presented later on. Note that the prior may be partially improper if the precision matrix $\boldsymbol{K}_j(\boldsymbol{\theta}_j)$ is not of full rank.

We complete the Bayesian specification by assuming inverse gamma prior for the error variance:
$$\sigma^2 \sim \text{IG}(a_0, b_0). \tag{4}$$
Given the model specification, this implies that the full conditionals of the variance parameter is also inverse gamma with updated parameters. In contrast, the full conditionals for the regression coefficients $\boldsymbol{\beta}_j$ are not available in closed form since unfortunately a normal prior in combination with an asymmetric normal observation models does not induce an asymmetric normal full conditional. We therefore construct proposal densities based on the penalised iteratively weighted least squares updates that would have to be performed to compute penalised expectile regression estimates in a frequentist backfitting procedure, i.e.
$$\hat{\boldsymbol{\beta}}_j^{[t+1]} = (\boldsymbol{B}_j^\top \boldsymbol{W}^{[t]} \boldsymbol{B}_j + \sigma^2 \boldsymbol{K}_j(\boldsymbol{\theta}_j))^{-1} \boldsymbol{B}_j^\top \boldsymbol{W}^{[t]} (\boldsymbol{y} - \boldsymbol{\eta}_{-j}^{[t]}),$$
where $\boldsymbol{B}_j$ is the design matrix associated with the j-th model term, $\boldsymbol{y}$ is the vector of responses, $\boldsymbol{\eta}_{-j} = \boldsymbol{\eta} - \boldsymbol{B}_j\boldsymbol{\beta}_j$ is the complete predictor without

the jth component and $\boldsymbol{W} = \text{diag}(w(y_1, \eta_1), \ldots, w(y_n, \eta_n))$. More precisely, we propose a new state for $\boldsymbol{\beta}_j$ from the normal distribution $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with expectation and covariance matrix given by

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j \boldsymbol{B}_j^\top \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\eta}_{-j}) \quad \text{and} \quad \boldsymbol{\Sigma}_j = (\boldsymbol{B}_j^\top \boldsymbol{W} \boldsymbol{B}_j + \sigma^2 \boldsymbol{K}_j(\boldsymbol{\theta}_j))^{-1}.$$

This framework covers, among others, individual-specific random effects, interaction surfaces based on either radial basis functions or tensor product splines, and varying coefficient terms as special cases and therefore provides a convenient generalisation of additive (mixed) models, see Fahrmeir et al. (2004). In this paper we will focus on linear effects, regularised by a spike and slab prior, penalised splines and Markov random fields, as those will be used in the analysis of childhood malnutrition. For each of those effects we will present an appropriate design matrix $\boldsymbol{B}_j$, which renders possible to estimate the predictor $\boldsymbol{\eta}$ as the sum over products $\boldsymbol{B}_j \boldsymbol{\beta}_j$ in the above setting. We will furthermore explain the corresponding penalisation/regularisation matrices and the resulting full conditionals.

2.3.1 Regularization of linear effects

We have chosen to present a spike and slab approach with the prior on the variances (see) as an option for regularisation. There are a lot of different ways to approach this problem, and changing the setup to other spike and slab approaches or the LASSO should be straight forward. In this case however, the design matrix $\boldsymbol{B}_j = \boldsymbol{X}$ simply is the data matrix. The regularisation matrix $\boldsymbol{K}_j(\boldsymbol{\theta}_j)$ is a diagonal matrix with entries $\frac{1}{\tau_j}$, where the τ_j are sampled in the same way as for spike and slab regularisation in mean regression. The τ_j themselves are assigned a mixture distribution as prior that comprises a slab part, being flat and spike part shifting the parameter towards zero: $\tau_j^2|\nu_j \sim (1-\nu_j)\text{IG}(a_{\tau^2}, \nu_0 b_{\tau^2}) + \nu_j \text{IG}(a_{\tau^2}, b_{\tau^2})$. This is realized by the additional parameter ν_0 in the inverse gamma distribution. If this parameter is chosen to be sufficiently small the form of the corresponding inverse gamma distribution has the required spike shape. The component of the mixture distribution of which τ_j is sampled, is controlled by the additional parameter ν_j which is assigned a Bernoulli prior $\nu_j \sim \text{B}(1, \theta)$ with a Beta hyperprior $\theta \sim \text{Beta}(a_\theta, b_\theta)$. The latter can be either chosen to be non informative or can be used to obtain an especially low or high number of parameters in the model. The resulting full conditionals for those two distributions are due to conjugacy $\nu|\cdot \sim B(1, \theta^*)$ with parameter

$\theta^* = \frac{\theta \text{IG}(\tau_j^2, a, b)}{\theta \text{IG}(\tau_j^2, a, b) + (1-\theta)\text{IG}(\tau_j^2, a, \nu_0 b)}$ and $\theta|\cdot \sim \text{Beta}(a_\theta + \sum_{j=1}^k \nu_j, b_\theta + k - \sum_{j=1}^k \nu_j)$. The full conditional for the τ_j is again a mixture of two inverse gamma distributions: $\tau_j^2|\cdot \sim (1-\nu_j)\text{IG}(a_{\tau^2}+0.5, \nu_0 b_{\tau^2}+0.5\beta_j^2)+\nu_j \text{IG}(a_{\tau^2}+0.5, b_{\tau^2}+0.5\beta_j^2)$.

2.3.2 Continuous Effects

We model the continuous non linear variables by penalised splines, see Eilers and Marx (1996) for details. The Bayesian formulation (see Brezger and Lang, 2006) requires the design matrix $\boldsymbol{B}_j$ to contain the basis functions for the B-splines and the precision matrix to operate as smoothing matrix. Therefore $\boldsymbol{K}_j(\boldsymbol{\theta}_j) = \frac{1}{\delta_j^2}\boldsymbol{D}_k^\top \boldsymbol{D}_k$ with $\boldsymbol{D}_k$ being the matrix of differences of kth order and δ_j^2 the smoothing parameter. The smoothing variance δ_j^2 is assigned an inverse gamma distribution:

$$\delta_j^2 \sim \text{IG}(a_j, b_j), \tag{5}$$

which, just as for the model variance, leads to an inverse gamma full conditional.

2.3.3 Spatial Effects

For the spatial effects we incorporate Markov random fields in our model. The design matrix $\boldsymbol{B}_j$ consists of the indicator function for the regions and the precision matrix $\boldsymbol{K}_j(\boldsymbol{\theta}_j) = \frac{1}{\delta_j^2}\boldsymbol{K}_j$, where $\boldsymbol{K}_j$ is the neighboring or adjacency matrix and δ_j^2 the smoothing parameter. The latter, just as for the continuous effects, is assigned an inverse gamma prior and also results to have an inverse gamma full conditional.

3 Simulations

Since we need a misspecified likelihood for our estimations, we aim to show that the resulting estimated expectiles are nevertheless valuable. We therefore conduct simulation studies comparing Bayesian expectile estimates with least squares and boosting estimates (Sobotka and Kneib, 2012) in order to quantify the performance of the procedures. The estimates are rated according to the true expectiles of the error distribution.

First, we evaluate the posterior mean as a point estimate and afterwards we explore coverage rates and the widths of the credible intervals.

3.1 Point Estimates

3.1.1 Design

To start the evaluation of the Bayesian expectiles and for comparison with existing alternatives, we generate two covariates, $X_1 \sim B(1, 0.5)$ and $Z_2 \sim U(0, 3)$ in sample sizes of $n = 100, 500$. Next, the random errors ε are drawn from an A) $N(0, 0.5z_2^2)$, B) $\text{Exp}\left(\frac{1}{z_2}\right)$ or C) $t(2)$ distribution. Together they comprise data for two simple semiparametric models in the following way:

$$\begin{aligned} \text{(M1)} \quad \boldsymbol{y} &= 2\boldsymbol{x}_1 + 5\exp(-0.5\boldsymbol{z}_2^2) + \boldsymbol{\varepsilon} \\ \text{(M2)} \quad \boldsymbol{y} &= 2\boldsymbol{x}_1 + 5\sin(2\boldsymbol{z}_2) + \boldsymbol{\varepsilon}. \end{aligned}$$

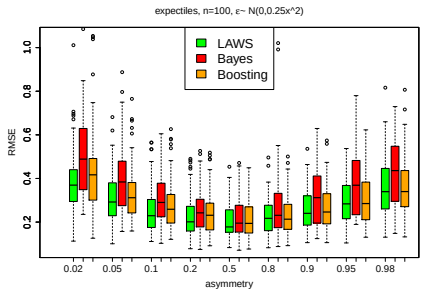
Hence, we have a challenging homoscedastic scenario (C) with infinite variance and two heteroscedastic scenarios, one of them with skewed errors (B). For each of the combinations of sample size, error distribution and model formula we generate 100 replications. The data are then analysed for expectiles with asymmetries $\tau \in \{0.02, 0.05, 0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 0.98\}$. We estimate the Bayesian expectiles with overall 35000 MCMC iterations, where 5000 are burn-in and we use a thinning of 30. This leaves us with a sample of 1000 observations from the posterior. This method is compared with a least asymmetrically weighted squares (LAWS) estimate and an estimate obtained with the use of component-wise functional gradient boosting, both as presented in Sobotka and Kneib (2012). The smoothing parameter in LAWS estimation is optimised with an asymmetric cross-validation criterion, for boosting the optimal stopping iteration from 1000 initial boosting iterations is also determined via cross-validation. For all algorithms, the nonlinear effect is estimated using a cubic B-spline basis with 20 inner knots and second order difference penalty. The methods are taken from the software package expectreg (Sobotka et al., 2013) available for R (R Development Core Team, 2013).

3.1.2 Results

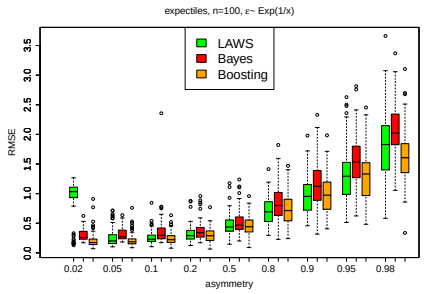
The quality of the results will be measured in terms of a root mean squared error for the estimated function:

$$\text{RMSE}(f_\tau) = \sqrt{(\boldsymbol{f}_\tau - \hat{\boldsymbol{f}}_\tau)'(\boldsymbol{f}_\tau - \hat{\boldsymbol{f}}_\tau)}.$$

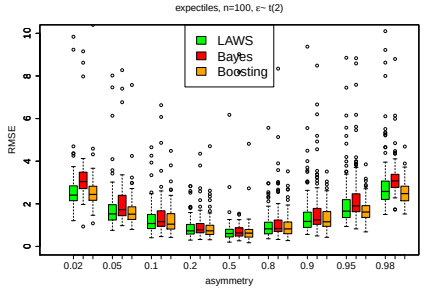
In Figure 1 we present the results of the three methods, for each expectile in direct comparison. The results are shown for $n = 500$ and exemplary for (M1). The complete results are available as online supplement. Our simulations show that, in terms of RMSE, the methods are quite interchangeable. The posterior means offer the same estimation quality as LAWS and boosting, at least within $\tau \in [0.05, 0.95]$. For more extreme expectiles, it seems that the numerical errors in the MCMC start to become more substantial, i.e. it becomes difficult to draw from the respective multivariate normal distribution with extreme weights. Otherwise, the choice for one of the estimates can be made regarding the outer properties, practicability or just personal habit, now that a Bayesian estimate is available. The choice might also depend on interval estimates rather than point estimates. The former are analysed in the next part of the simulations.



(a) $N(0, 0.25z_2^2)$-error



(b) $\text{Exp}(\frac{1}{z_2})$-error



(c) $t(2)$-error

Figure 1: RMSE for $n = 100$, the three different errors, separately for each expectile. Boxplots created from 100 replications.

3.2 Interval Estimates

3.2.1 Design

Similar as in Waldmann et al. (2013) we compare 95% credible intervals with frequentist confidence intervals based on an asymptotic normal distribution (Sobotka et al., 2013). Confidence intervals from boosting would be obtained with a computationally demanding nonparametric bootstrap and are therefore omitted. The comparison is made regarding coverage properties and interval widths. For a simpler visualisation we focus on a single nonlinear effect:

$$\text{(M3)}\ \boldsymbol{y} = \sin(2(4\boldsymbol{z} - 2)) + 2\exp(-16^2(\boldsymbol{z} - 0.5)^2) + \boldsymbol{\varepsilon}.$$

The covariate is drawn from a $U(0, 1)$ distribution, the error from a $N(0, (0.2+|z - 0.5|)^2)$ and an $\text{Exp}\left(\frac{1}{0.2+|z-0.5|}\right)$ distribution. The nonlinear effect then has its highest frequency as well as lowest variance at 0.5 while the variance increases with$z \to 0$ and $z \to 1$. The frequentist asymptotics start to apply from 500 observations and for extreme expectiles, 1000 observations are recommended. Hence, we generate data sets with $n = 500, 1000$ and in order to properly measure the coverage rate, we generate 1000 replications. The rest of the parameters remain as before.

3.2.2 Results

We measure the coverage of the confidence intervals at a given covariate value z_i as

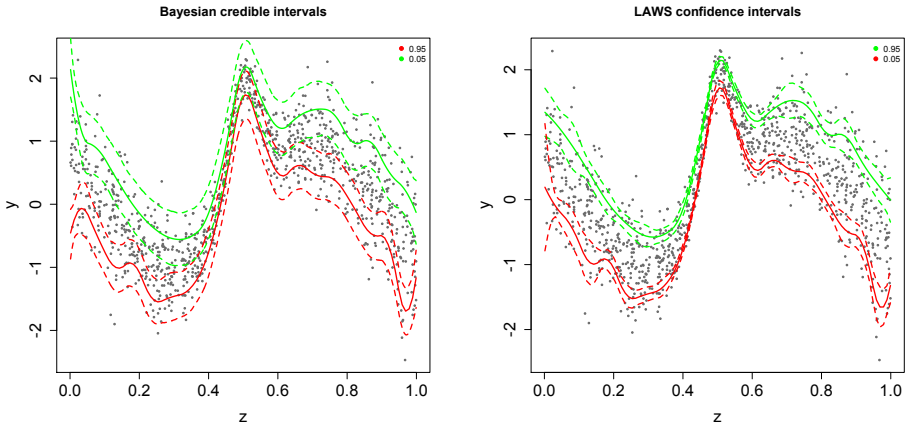
$$\widehat{\text{Cover}}(CI(\hat{f}_{j,\tau}(z_i)) = \frac{1}{1000}\sum_{k=1}^{1000} \mathbb{1}_{\{\hat{f}_{j,\tau}(z_i) \in CI(\hat{f}^{[k]}_{j,\tau}(z_i))\}},$$

the maximum width of all confidence intervals at all fixed z_i

$$\max \widehat{\text{Width}}(CI(\hat{f}_{j,\tau}(z_i))) = \max_k(\hat{f}^{[k]}_{j,\tau,U}(z_i) - \hat{f}^{[k]}_{j,\tau,L}(z_i))$$

where f_U and f_L denote the upper and lower ends of the interval estimate. The minimum width is determined in the same way. The evaluations are done on a regular grid of length 100 within the covariate domain. In Figure 3 we can see that the coverage of both interval estimates is rather poor at the center of the covariate where the curvature of the generating function is high. This might result from a bias that comes with the addition of the penalty. Otherwise the plots show that the width of the frequentist intervals

generally increases with increasing variance in the errors while the credibility intervals remain at the same width over the whole covariate domain. The effect is a better coverage at the center of the covariate and worse coverage for strongly increasing variance regions. This result is especially visible in Figure 2 where two estimates and intervals are shown for an exemplary data set. Here we can see that the confidence intervals are much narrower in the center than the credible intervals. Reasons for this behaviour can be found in the misspecified likelihood which is just an auxiliary tool to fit the point estimates and does not describe the data well. Hence, the estimated variance of the fit is constant. The results for exponential errors and for a sample size of 500 are available as online supplement.



(a) Bayesian example analysis (b) frequentist example analysis

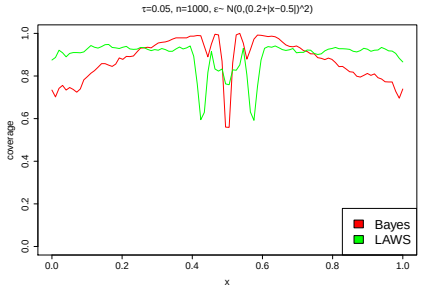
Figure 2: Examplary estimates and pointwise intervals for $n = 1000$, (M3) and normal errors obtained from MCMC and LAWS estimation.

3.3 Regularisation

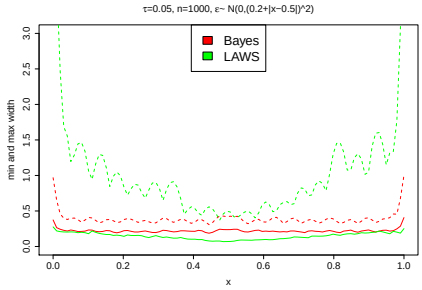
3.3.1 Design

We choose a simulation setup close to the one provided in the original LASSO paper by Tibshirani (1994). The model setup is a simple linear regression with

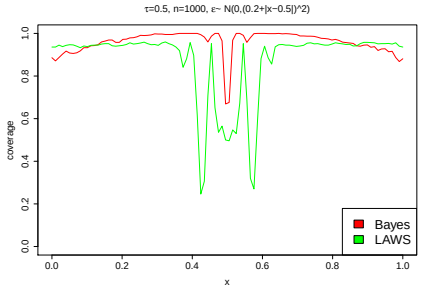
$$\text{(M4)}\ \boldsymbol{y} = 3.3 + 1.5\boldsymbol{x}_1 + 0\boldsymbol{x}_2 + 0\boldsymbol{x}_3 + 2\boldsymbol{x}_4 + 0\boldsymbol{x}_5 + 0\boldsymbol{x}_6 + \boldsymbol{\varepsilon}.$$



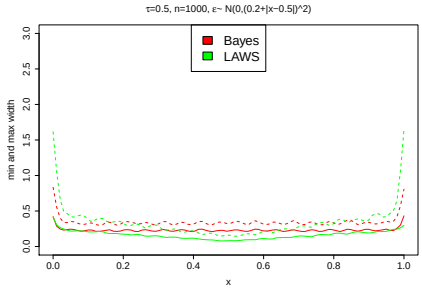
(a) coverage for $\tau = 0.05$



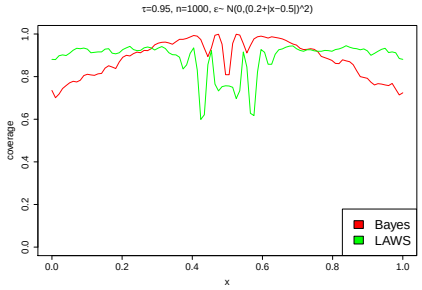
(b) min and max width for $\tau = 0.05$



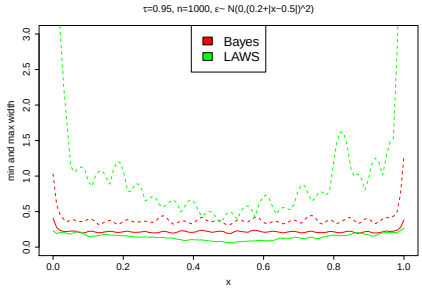
(c) coverage for $\tau = 0.5$



(d) min and max width for $\tau = 0.5$



(e) coverage for $\tau = 0.95$



(f) min and max width for $\tau = 0.95$

Figure 3: Coverage rates on the left and interval widths on the right for normal errors and $n = 1000$. Minimal interval width given in solid, maximum width in dashed lines.

Thus we have only three covariates, that really have an effect, whereas the rest is zero. The x_k are drawn from a multivariate normal distribution with mean 0, standard deviation 1 and pairwise correlation $Cor(x_k, x_l) = 0.5^{|k-l|}$.

Further we have $\varepsilon \sim N(0,1)$. We generate datasets of three different sizes, $n = 20$, $n = 50$ and $n = 100$ and perform 50 replications for each scenario. In every replication the estimated regression coefficients are recorded and for each covariate the acceptance probability for inclusion into the model.

3.3.2 Results

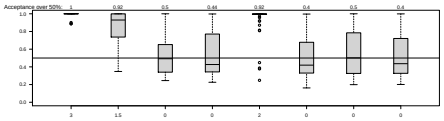
We report the results exemplary for $n = 50$ since we find no substantial differences in the simulation results for different sample sizes. As can be seen in Figure 4, the results are generally quite convincing. The left panel shows the boxplots for the percentage of simulations in which the acceptance probability for the corresponding parameter was over 0.5. The plots in the right column display the estimation of the regression parameters β_k themselves. Note the fact that in the left panel there is one more box than on the right side, due to the fact that the intercept was not subject to regularisation. These plots show that in general the regularisation makes more false positive errors as we have a quite high rate of correctly chosen real effects whereas part of the $\beta_k = 0$ get acceptance probabilities over 50% and are thus categorized incorrectly. In general the performance is better in the central expectiles than in the outer ones. However, if one does not use a hard 50% criterion but more of a comparison between the acceptance probabilities for the covariates the structure gets quite obvious over all different covariates.

3.4 Simulation roundup

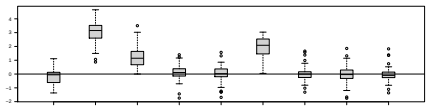
Overall we can say that boosting is a flexible tool that results in good point estimates, confidence intervals for large data sets with strong heteroscedasticity might be more reliable with a LAWS estimate, but the estimated Bayesian expectiles are as efficient and provide better coverage for small samples. The addition of Spike-and-Slab priors provides a powerful tool for model selection which should nevertheless be controlled manually.

4 Example

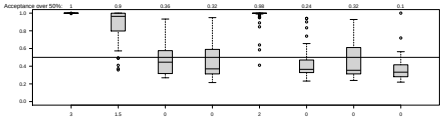
A data set consisting of 5389 observations of children was obtained from the Demographic Health Surveys (DHS, `www.measuredhs.com`). The study was conducted in Tanzania in 1992. It contains information on weight, height, sex



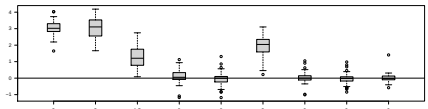
(a) acceptance probabilities for $\tau = 0.05$



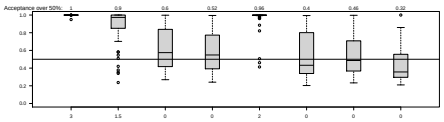
(b) estimations for β for $\tau = 0.05$



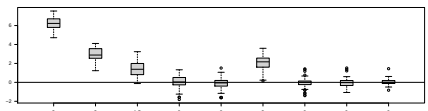
(c) acceptance probabilities for $\tau = 0.5$



(d) estimations for β for $\tau = 0.5$



(e) acceptance probabilities for $\tau = 0.95$



(f) estimations for β for $\tau = 0.95$

Figure 4: Boxplots for acceptance parameters being over 0.5 in the left panel, average of acceptance probabilities above each box. Boxplots of the estimations of the parameter β in the right. Note that β_0s for the different expectile levels are supposed to be different.

and age of the children themselves, information about the parents - namely the mother - such as her BMI at the birth of the child, her educational background (in four categories) as well as her current employment status (either employed or unemployed) and the information on the residence. The latter actually splits into two: the categorical variable on the surrounding (urban or rural) and the spatial variable, indicating the province mother and child are living in. Chronic malnutrition leads to *stunting* (insufficient height for age) which will be used as measure for the extent of undernourishment. The height of the children is compared to a reference population of supposedly healthy children of the same age in a so called *z-score*: $z_i = (AI_i - MAI)/\sigma$. In this formula AI_i stands for the stunting index of child i, MAI the median stunting value in the reference population and σ for the standard deviation of stunting in the reference population. The mean value in our data set is -177.9, the standard deviation 142.24, 90% of the children have a z-score lower than zero and the 95%-quantile reaches from -455.00 to 108.05. As explained in the introduction, we use a model incorporating the continuous covariates non-

linearly, the categorical variables linearly and the province as a spatial effect (see equation(3)). The estimation was executed as described in Section 2: the nonlinear effects are modeled with Bayesian P-splines, the spatial effect with a Markov random fields and to judge the importance of the categorical covariates in the different expectiles, we used spike and slab priors for regularisation. The model was estimated for the $\tau = 0.05, 0.1, 0.2, 0.5, 0.8, 0.9$ and 0.95-expectiles.

The results for the linear effects are displayed in Table 1. The proportion of MCMC iterations in which the effect was considered different from zero by the spike and slab selection is presented in the line below. Effects, for which this proportion exceeds 50% are printed in boldface. For reasons of clarity only five different expectiles are displayed, the rest behaves analogously. Note that for the covariate *work* the sign of the effect changes over the expectiles. This means that the impact of the employment of the mother is positive in the lower parts of the conditional distribution, whereas it has a negativ impact in the middle to higher ends. A similar effect can be seen for the impact of *secondary school* in comparison to *no education at all*. The positive effect of the variable *rural* is no surprise, as the proximity to the farms and the traditional higher family bonding is of high importance for the adequate supply. The negative effect of sex simply displays the fact, that boys of this age are generaly less tall than girls.

Nonlinear effects are displayed in Figure 5. The results are very close to those from Kandala et al. (2001), where the data set was analysed in a mean regression setting. There are small differences between the expectiles, but in general the effects are stable over the whole distribution.

For the spatial effects see Figure 6. The effect of the capital Dar es Salaam in the east of the country is positive over all expectiles, which is in contradiction to the negative effect of the variable *urban* in comparison to *rural*. Thus we conclude, that the effect of the capital as being better supplied than the rest of the country voids this effect. Another fact worth mentioning is the positive effect of the south west on the higher expectile. This region neighbours Lake Tanganyika and is known for its fertility.

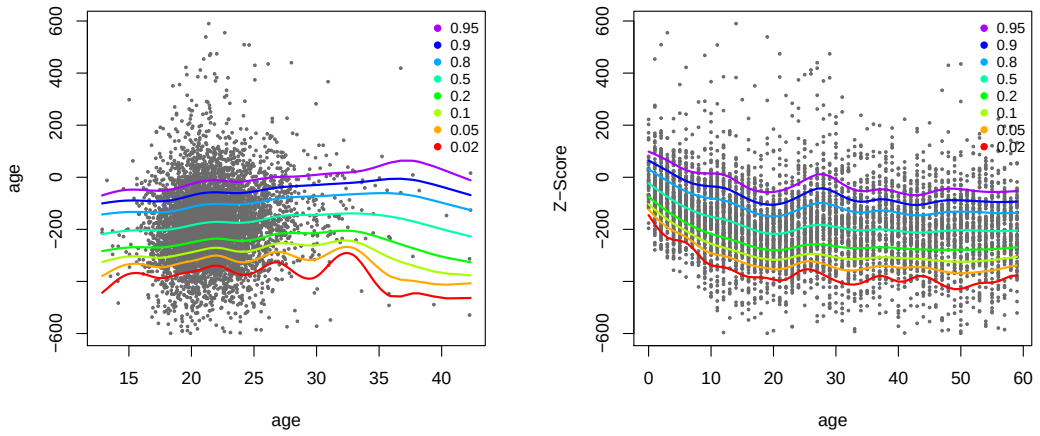
5 Conclusion

The Bayesian formulation of expectile regression outlined in this paper provides both the Bayesian counterpart to frequentist expectile regression and

Variable / τ	0.05	0.2	0.8	0.95
mother's work	**4.15**	0.14	**-5.45**	**-13.62**
reference: "unemployed"	*0.96*	*0.26*	*0.99*	*1.00*
mother's education:	*reference: "no education"*			
"primary school"	**-1.93**	-0.21	**-7.55**	**-17.66**
	0.52	*0.34*	*0.87*	*0.99*
" secondary school"	**19.79**	**14.64**	0.56	**-7.80**
	1.00	*1.00*	*0.36*	*0.84*
"higher education"	**61.62**	**55.60**	**59.67**	**78.22**
	1.00	*1.00*	*1.00*	*1.00*
mother's residence	**12.56**	**13.73**	**10.63**	**3.10**
reference: "urban"	*1.00*	*1.00*	*1.00*	*0.69*
child's sex	**-5.27**	**-5.62**	**-3.52**	**-1.95**
reference: "female"	*0.99*	*1.00*	*0.94*	*0.68*

Table 1: Estimated parametric effects for Childhood Malnutrition data. Reference categories and spike and slab acceptance probabilities obtained by MCMC are included in italics. Covariates accepted more often than 50% of the time are set in boldface.

the expectile analogue to Bayesian quantile regression. While standard semiparametric regression specifications in expectile regression can already be handled in a frequentist setting based on iteratively weighted least squares estimation, the Bayesian formulation opens up the possibility to include more complex regression specifications. We showed this by applying Spike-and-Slab regularisation to expectile regression which is to our knowledge the first automatized variable selection approach implemented in the expectile regression context. Further extensions could be the Dirichlet process mixture priors for random effects or Bayesian regularisation priors using a conditional Gaussian prior structure as suggested for Bayesian quantile regression in Waldmann et al. (2013). Moreover, Bayesian expectile regression comprises the determination of the smoothing variances δ_j^2 as an integral part of the inferential procedure and provides measures of uncertainty also for complex functionals of the model parameters. However, the asymmetric normal likelihood will usually induce a model misspecification and the impact of this misspecification will have to be studied in detail.

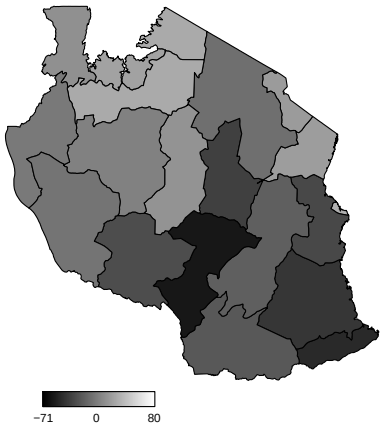


(a) Nonlinear effect for BMI of mother at birth (b) Nonlinear effect for the age of the child in months

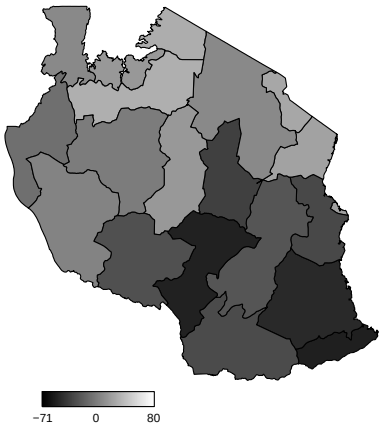
Figure 5: Estimated nonlinear effects for the childhood malnutrition data. Results for expectiles from 0.05 to 0.95 shown.

A further integral part of this misspecification can also be found in the interval estimates constructed from the MCMC algorithm. These fail in terms of coverage for large samples and strong heteroscedasticity while the quality of the point estimates proves satisfying. That is at least in comparison to a "classical" LAWS estimate, for example. In consequence, the overall questions about expectile regression remain unchanged and independent from the estimation procedure.

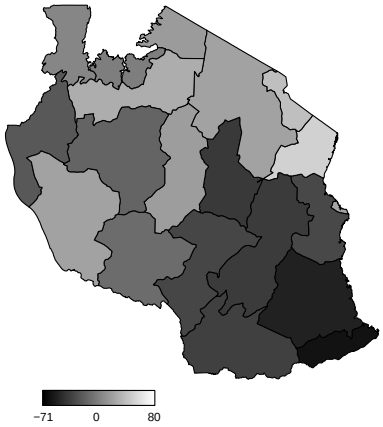
Two of the main questions regarding expectile regression are the crossing of expectile curves and the interpretation of single expectiles. While non-crossing estimates exist in a frequentist setting and have been proposed in different complexity by Sobotka and Kneib (2012) and Schnabel and Eilers (2012), it would be at least challenging to apply them in a boosting or Bayesian setting. Regarding the interpretation of the estimates, additional arguments to the ones presented in Section 2 are presented by Schulze Waltrup et al. (2013). However, both questions remain in the focus of research regarding expectiles.



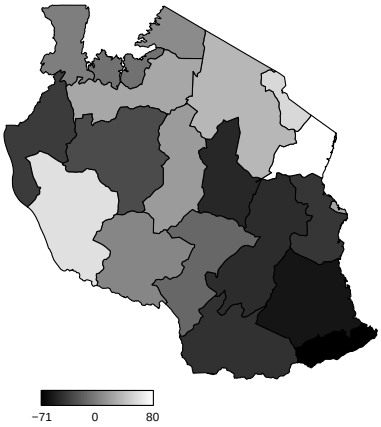
(a) 0.05-expectile



(b) 0.2-expectile



(c) 0.8-expectile



(d) 0.95-expectile

Figure 6: Estimated spatial effects for the childhood malnutrition data provided in a map of Tanzania.

References

Brezger, A. and S. Lang (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis 50*, 967–

991.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica 1*, 93–125.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing using B-splines and penalized likelihood. *Statistical Science 11*, 89–121.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica 14*, 731–761.

Fenske, N., T. Kneib, and T. Hothorn (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association 106*(494), 494–510.

Kandala, N. B., S. Lang, S. Klasen, and L. Fahrmeir (2001). Semiparametric analysis of the socio-demographic and spatial determinants of undernutrition in two african countries.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika 81*, 673–680.

Kozumi, H. and G. Kobayashi (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation 81*, 1565–1578.

Lum, C. and A. Gelfand (2012). Spatial quantile multiple regression using the asymmetric laplace process. *Bayesian Analysis 7*(2), 235–258.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica 55*, 819–847.

Reed, C. and K. Yu (2009). A partially collapsed gibbs sampler for bayesian quantile regression. Technical report, Department of Mathematical Sciences, Brunel University.

Reich, B. J., H. D. Bondell, and H. Wang (2010). Flexible bayesian quantile regression for independent and clustered data. *Biostatistics 11*, 337–352.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org.

Schnabel, S. K. and P. Eilers (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis 53*, 4168–4177.

Schnabel, S. K. and P. Eilers (2012). Expectile sheets for joint estimation of expectile curves. *Statistical Modelling, under review*.

Schulze Waltrup, L., F. Sobotka, T. Kneib, and G. Kauermann (2013). Quantile or expectile regression - is there a favorite? *submitted*.

Sobotka, F., G. Kauermann, L. Schulze Waltrup, and T. Kneib (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing 23*(2), 135–148.

Sobotka, F. and T. Kneib (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis 56*, 755–767.

Sobotka, F., G. Marra, R. Radice, and T. Kneib (2013). Estimating the relationship between women's education and fertility in botswana by using an instrumental variable approach to semiparametric expectile regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(1), 25–45.

Sobotka, F., S. Schnabel, and L. Schulze Waltrup (2013). *expectreg: Expectile and Quantile Regression*. R package version 0.37.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics 6*, 231–252.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Waldmann, E., T. Kneib, Y. R. Yue, S. Lang, and C. Flexeder (2013). Bayesian semiparametric additive quantile regression. *Statistical Modelling, to appear*.

Yu, K. and R. A. Moyeed (2001). Bayesian quantile regression. *Statistics & Probability Letters 54*, 437–447.

Yue, Y. and H. Rue (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics and Data Analysis 55*, 84–96.

Ziegler, J. F. (2013). Coherence and elicitability. Technical report.